Psychological Monographs

General and Applied

No. 394
1955

Martha Littleton Kelly

Vol. 69
No. 9

A Study of Industrial Inspection by the Method of Paired Comparisons

By

Martha Littleton Kelly
Aurora College, Aurora, Illinois

Price $1.00

Edited by Herbert S. Conrad

Published by The American Psychological Association, Inc.

Psychological Monographs: General and Applied

Combining the *Applied Psychology Monographs* and the *Archives of Psychology* with the *Psychological Monographs*

MANUSCRIPTS should be sent to the Editor.

Because of lack of space, the *Psychological Monographs* can print only the original or advanced contribution of the author. *Background and bibliographic materials must, in general, be totally excluded* or kept to an irreducible minimum. Statistical tables should be used to present only the most important of the statistical data or evidence.

The first page of the manuscript should contain the title of the paper, the author's name, and his institutional connection (or his city of residence). Acknowledgments should be kept brief, and appear as a *footnote* on the first page. No table of contents need be included. For other directions or suggestions on the preparation of manuscripts, see: CONRAD, H. S. Preparation of manuscripts for publication as monographs. *J. Psychol.*, 1948, 26, 447-459.

CORRESPONDENCE CONCERNING BUSINESS MATTERS (such as author's fees, subscriptions and sales, change of address, etc.) should be addressed to the American Psychological Association, Inc., 1333 Sixteenth St. N.W., Washington 6, D.C. Address changes must arrive by the 25th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee third-class forwarding postage.

Vol. 69, No. 9 Whole No. 394, 1955

Psychological Monographs: General and Applied

A Study of Industrial Inspection by the Method of Paired Comparisons[1]

Martha Littleton Kelly
Aurora College, Aurora, Illinois

INSPECTION is an indispensable part of our industrial production system. Every manufactured product is examined for quality at least once—if only by the customer prior to purchase. Most industries, aware of the necessity for maintaining the quality of the goods they offer in a competitive market, subject their product to several inspections during manufacture. Few companies, however, have devoted to inspection a fraction of the attention given to improving the product or production process. They seem to consider inspection a necessary evil, an unproductive element of manufacturing cost which should be held to a minimum.

Statistical quality control probably owes much of its growing popularity to the promise it offers for reducing inspection costs. It concentrates on reducing the volume of product inspected and usually accepts the existing method; its formulas are based on the somewhat dubious assumption that any individual inspector's work will be 100% accurate (7).

Basically, the inspector must examine a unit of product and then decide whether it is as good as the quality specified for it. Psychological techniques should be particularly appropriate to investigation of the perception and judgment factors involved. But psychologists have been even less aware of the possibility of improving the inspection process than have management and production engineers, if the fact that only 16 studies on inspection appeared in psychological journals over a ten-year period is any indication of their interest. A majority of the published studies (2, 3, 4, 5, 8, 10, 14, 15, 16, 17, 18, 19) have investigated the relationship between one or more psychological or "aptitude" tests and a criterion of job performance. Not only are the validity coefficients reported too low to be generally useful, but most of the criteria were derived from data unrelated to the subject's ability to make correct decisions on the job. Four published work-sample experiments (1, 6, 13, 20) and two unpublished studies (9, 11) did report specific ability factors related to job performance (such as near-point visual acuity) and/or pertinent job factors (such as lighting). However, the relationships reported were not high enough nor were the studies in sufficient agreement with respect to their specific conclusions to be applicable to jobs other than the ones studied.

[1] Based on a thesis submitted to the faculty of Purdue University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, May, 1953. The writer wishes to express her appreciation to Dr. Joseph Tiffin, chairman, and the other members of her committee for their interest and guidance, and to the management of Corning Glass Works, Electrical Products Division, who supplied the material and job information on which this study was based.

The reader in search of information that he can use to improve the inspection job that *he* is concerned about finds all this interesting, but somewhat frustrating. If he makes a critical analysis, he notes what may be a significant omission in all of the studies. None of them attempted either description or analysis of the specification[2] controlling the inspection process, so there is no point of departure for comparison of the results of the separate investigations. Individual differences in the subjects' knowledge of the specification were unknown; hence that factor had to be treated as an uncontrolled variable. One study assumed that this job knowledge may depend on experience, but little relationship between experience and performance was found (20).

It may then appear to the person in search of usable information that if inspection could be studied by a method focused sharply on the actual judgments involved, with differences in job knowledge either controlled or eliminated as a variable, he might find his data indicative of the basic inspection function and his conclusions applicable to different inspection jobs. His problem then becomes one of experimental design.

A work sample designed by the method of paired comparisons[3] appeared particularly suited to the approach outlined above. Material from a regular job presented in this design would require a decision by the subject only as to which is the better piece in every pair that is presented for judgment. The data would be dependent on the judgments required by the job, but would not depend on knowledge of the specification. Inclusion of both good and reject pieces in the sample would give data that could be scored for accuracy. Analysis of the degree of differentiation made between pieces within the sample, together with the subjects' accuracy in judging good pieces as better than reject pieces, should indicate the adequacy of the present specification or be a guide to its revision. That is the background of the present study.

Inspection for appearance of television face plates or panels, a job familiar to the experimenter, was selected for study. The glass panels and job information were supplied by Corning Glass Works, Corning, New York. The specific methodology was worked out in the laboratory at Purdue University, and the investigation was completed at Corning's tube plant at Albion, Michigan.

[2] The formal statement or description of quality is called the specification. It may be stated in terms of measurable dimensions or tolerances, describe the appearance of a minimally acceptable unit of product, or be a sample of the minimally acceptable product. Knowledge of the specification, then, is the equivalent of job knowledge.

[3] Also called the "method of variable stimuli," this method was used by early psychologists in research on discrimination and judgment. A number of stimuli (their real value may be unknown) are presented in pairs, each stimulus paired with every other stimulus, for a judgment as to which of the pair seems larger, smaller, more pleasing, etc. to the subject. Each stimulus is given a relative value or *rank* based on the proportion of times it is preferred. This is an effective technique for measuring esthetic judgments, and it seemed suited to appearance inspection where the inspector decides simply on a basis of "how it looks to him."

THE PILOT STUDY

Since there seemed to be no precedent in the literature for a study of inspection by the method of paired comparisons, the need to develop the design empirically was obvious. It was equally evident that a frankly experimental ap-

proach would not be suitable for an industrial setting, so the pilot study was designed for the laboratory at Purdue University.

The preliminary investigation was planned to answer these questions: When an individual subject is presented a series of pairs of a unit of product, can he distinguish between the two pieces of a pair sufficiently to decide which is the better one? Will his judgments on the separate pairs reflect a consistent basis of distinction such that a summation of the times he prefers each piece will result in a *ranking* of all the pieces in the order of his preference? Will he make a similar ranking on a second trial? Will a group of subjects agree well enough on their rankings so that the mean rank per piece can be employed as a quality scale for the sample? Will a scale developed in this way be reliable?

The First Method

Procedure

The laboratory was set up to duplicate as closely as possible the actual workplace of the television panel inspector in the factory. A bench the height and width of the conveyor belt, together with a model of the inspector's booth to be rolled along the bench, was constructed from drawings supplied by the company. Two work samples of 10 panels each were selected by the experimenter from a 200-piece lot of 12½-inch round panels preinspected at the factory. Sample I consisted of 10 panels called good by the plant inspectors, but showing slight defects judged on appearance[4] only. Sample II consisted of 10 panels classified as rejects, showing the same types of defect. The design of presentation met the requirements of the paired-comparison method for random order of pairs and balanced placement of each piece on the right or left of the pair, within limits set by the size of the panels. The 10 panels were placed in order along the bench and the subject was instructed to judge the pairs 1–2, 3–4, etc. As he finished 9–10, the first panel was removed and he returned to the head of the bench to judge 2–3, 4–5, etc. As soon as he finished the pair 8–9, all pieces were removed and presented in an entirely new order. He judged them in pairs as before, making two trips along the bench to make nine comparisons. Four major shifts were required for the complete design.

Subjects were 11 staff members and graduate students in industrial psychology at Purdue University. Each subject made two trials on each sample, giving 45 judgments per sample or 180 judgments in all. Standard instructions to "examine each pair of panels and indicate which one would give a better picture in a television set" were read by the experimenter, who also presented the material and recorded the decisions. The number of times each subject preferred each panel was tabulated to get individual rankings or scale values per panel. Reliability of the individual rankings on each sample for two trials was calculated. Agreement among the subjects was checked by correlating the mean scale values assigned by each subject with the mean scale values assigned by all other subjects. Scale values assigned by individual subjects were averaged to derive mean scale values for each sample for the first trial, then for the second, and the reliability of the sample was computed by correlating the mean scale values of the two trials.

[4] The specification for these defects is contained in a minimally acceptable panel, called a *limit sample*.

Results

The results at this stage were encouraging. After they had examined the first two or three pairs, the subjects expressed little difficulty in reaching a decision. On the reasoning that if each subject were deciding merely on a chance basis every panel would tend to receive the same number of "votes," whereas if all his choices were clear cut, each panel would be given a different number of votes and the series would be ranked 1 to 10, the number of ranks obtained from each subject's preferences was noted (see Table 1).

The reader will recall that in the paired-comparison design, each piece is paired with every other one. In the 10-piece sample, each piece appears in nine pairs—is judged nine times. A piece consistently preferred to all the others receives nine votes. A piece preferred to all but one will get eight, and so on, the

TABLE 1

NUMBER OF RANKS RESULTING FROM EACH SUBJECT'S PREFERENCES (SAMPLES I AND II) AND RELIABILITY OF RANKS, PILOT STUDY)

Subject ($N = 11$)	Number of Ranks				Reliability, Two Trials	
	Sample I		Sample II		Sample I (rho)	Sample II (rho)
	Trial 1	Trial 2	Trial 1	Trial 2		
Ts	7	9	9	1	.85	.70
Ga	8	6	7	6	.76	.76
Ti	9	9	10	9	.76	.89
Ma	9	10	10	10	.92	.89
As	9	10	10	9	.95	.94
Ac	8	7	10	7	.87	.78
Wi	7	7	9	9	.80	.47
Ni	8	8	9	9	.93	.74
Co	9	9	10	8	.95	.78
McG	9	9	10	8	.76	.78
Fa	9	9	10	9	.90	.90
Mean, two trials	7.9		8.7		.93	.88
Median, two trials	8.7		9.2			
r'*					.80	.76

* Using paired scores of individual subjects, Trial 1 and Trial 2.

least preferred piece getting no votes for a score of zero. The range from nine to zero gives a maximum of 10 ranks. If there is no consistency in the judgments, each piece is as likely to be rejected as it is to be preferred. In this case, every piece would receive four or five votes, and there would be a spread of only two ranks.

In the pilot study, the minimum number of ranks was 6, the mean number being 7.9 ranks for Sample I and 8.7 ranks for Sample II. The median number of ranks was 8.7 for Sample I and 9.2 for Sample II. This indicates that each subject set up some basis for his decisions which differed considerably from chance.

All but one of the 22 rank-order reliability coefficients calculated for the indivudual scale values (or average ranks) were significantly higher than zero at the .01 level of confidence (see Table 1). For the two trials, they ranged from .76 to .95, averaging .88, by Fisher's z' conversion for Sample I. On Sample II they ranged from .47 to .94, averaging .81. (The Pearsonian r for all subjects together was .80 for Sample I and .76 for Sample II.) Rank-order correlation of the scale values assigned by individual subjects with the mean scale values assigned by all other subjects ranged from .54 to .97, averaging .93 for Sample I, .88 for Sample II, and .90 for both samples. Reliability of the mean scale values assigned by all subjects, shown in Table 2, was .96 for Sample I and .98 for Sample II. Although an exact test of the significance of the difference between these statistics is not available, the two samples appear to have similar reliability.

REVISED METHOD

Procedure

These data indicated satisfactory answers to the preliminary questions. However, the time required for administration was excessive. Accordingly, the

TABLE 2

MEAN SCALE VALUES (NUMBER OF FIRST CHOICES PER PANEL) AND RELIABILITY OF MEAN SCALE VALUES, SAMPLES I AND II, PILOT STUDY

Sample and Trial	Panel No.									
	1	2	3	4	5	6	7	8	9	10
Sample I										
Trial 1	6.9	6.5	4.9	3.0	1.6	5.8	1.8	6.2	2.1	6.2
Trial 2	7.2	7.0	4.9	3.4	1.2	5.8	1.7	5.5	2.5	5.6
Both trials	7.1	6.8	4.9	3.2	1.4	5.8	1.8	5.9	2.3	5.9
Sample II										
Trial 1	4.4	3.9	5.9	.73	4.6	6.0	3.0	2.8	8.2	5.5
Trial 2	4.9	3.6	5.8	1.2	4.4	6.5	3.2	1.9	7.4	6.0
Both trials	4.7	3.8	5.8	.9	4.5	6.3	3.1	2.3	7.8	5.8

order was "streamlined" to minimize handling of the panels, with some sacrifice of the precise control of time and place association characteristic of the first order of presentation. In this order, the subject judged the pairs 1–2, 2–3, 3–4, etc., giving nine judgments on a single trip down the bench. As he finished the pair in positions 9–10, the panels in the odd-numbered positions were shoved to the rear of the bench. Those which had been in the even-numbered positions were shoved up, in order, to positions 1–5, while those which had been in the odd-numbered positions were shoved down the bench to positions 6–10. The subject compared adjacent pieces as before. The panels were again shifted as described above. Four such shifts gave the required 45 pairs.

A new group of 17 students and staff members repeated the experiment in the new design.

A further test of the stability of the scale values was made by combining five panels from each sample into a third work sample which was presented to the new group of subjects.

Results

The mean scale values assigned by the second group of subjects correlated with those assigned by the first group at .97 for Sample I and .95 for Sample II, indicating that they were not affected by the change in design. These data are shown in Table 3.

When five panels from Sample I were combined with five from Sample II to make Sample III, it was not expected

TABLE 3

MEAN SCALE VALUES

SECOND GROUP OF SUBJECTS, PILOT STUDY

Panel No.	Sample I	Sample III
1	7.73	
2	7.00	6.97
3	4.81	5.06
4	3.50	
5	1.00	
6	5.77	6.53
7	.73	
8	6.58	
9	2.31	3.00
10	5.22	6.18
	Sample II	
1	4.08	
2	4.58	
3	5.50	4.24
4	1.42	.36
5	3.67	2.76
6	7.17	
7	2.58	1.46
8	2.42	
9	7.42	8.03
10	6.00	

that the scale values of individual panels would remain the same. But an order of preference for each set of five similar to their order in the original samples would indicate that distinctions actually were being made on the basis of pairs rather than being determined by the over-all composition of the sample. Both sets of five were ranked in the same order in the new sample. Moreover, the amount of difference in scale value between adjacent panels remained remarkably similar, as shown in Table 3 and Fig. 1.

These data indicated that untrained subjects, when comparing two pieces of product according to the method of the experiment, were able to make consistent distinctions between the pieces. The preference scale developed from their judgments was reliable and consistent for a second group of subjects. The preference rank of groups of panels remained the same when these groups were incorporated in another sample.

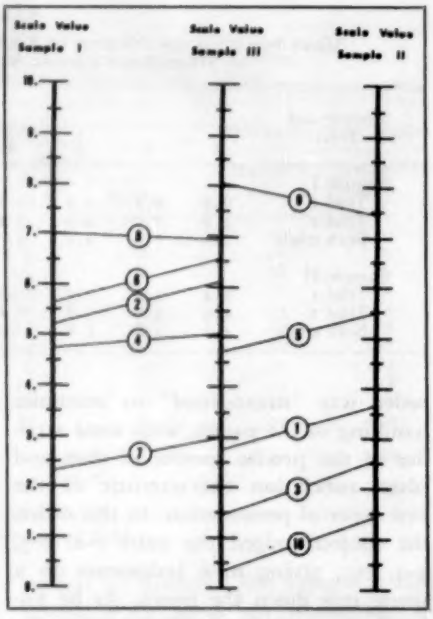

FIG. 1. Original scale values of 5 panels each from Samples I and II, with new scale values assigned them in Sample III. (Number in circle indicates panel number in Sample III.)

THE FACTORY STUDY

Presumably, industrial subjects have been trained to judge units of product as good or bad, the basis of their judgments being defined by the specification. If they were required to develop their judgments from the material inspected, as the laboratory subjects had done, would their decisions result in a similar ranking of the pieces in the sample, or would they rank all good pieces as equal and give equivalent ranks to all rejects? The factory study was set up to answer this question as well as to get information on the accuracy of performance in distinguishing acceptable from reject units of product.

It was expected that the panels scaled in the laboratory would be used, but by the time the pilot study had been completed, the 12½-inch panel was obsolete. Consequently, a new sample from current production was developed in the plant with the assistance of the Customer Inspector.[5] Otherwise, the methodology was identical to that of the pilot study.

[5] This man is the company's final authority on inspection of individual items. A member of the Sales Department, he works with the product engineers of the company and customer when the specifications are set up. When a customer complains of items in a shipment that do not meet the "specs," he inspects all contested pieces at the customer's factory. His judgment is accepted as final and credits or replacements are made only at his direction.

Procedure

The Work Sample

The sample consisted of ten 20-inch rectangular television panels selected by the Customer Inspector from a pool of approximately 40 set aside from current production by the Quality Control Department. Only panels containing defects of appearance (compared to an official limit sample[6]) were included. The inspector tried several assortments, the final one yielding a rank-order reliability of .94 ($r = .90$) for two trials. It included four rejects and six good panels, one of which was the equivalent of the limit sample for the defects represented.

The Paired-Comparison Trials

The experiment was set up near the production line, using the regular equipment for special inspection jobs—a long workbench and a portable inspector's light which could be rolled along it. The sample was presented in the "streamlined" order to 12 quality control and 2 line inspectors. They were instructed only to "look at a pair of panels and then tell the experimenter which of the pair is better." If they asked questions, they were told not to worry about the specifications, simply to decide on the two panels being considered. Two supervisors also participated.

The Job-Practice Run

Following completion of the paired comparisons, each subject was asked to inspect the 10 panels individually just as he would on the job. His decision to pass, reject, or send each panel for reclaiming was recorded. The following description of job method seems pertinent here.

Inspection method. Most inspectors work on either side of a slowly moving conveyor belt which carries the panels or completed bulbs out of the controlled cooling oven or lehr. The inspector places the good pieces in cartons and tosses the rejects into a hopper. It is common, but not invariable, practice to require each inspector to keep a record of the pieces rejected, noting the defect for which each has been discarded.

The line inspectors are usually new hires, while the quality control inspectors are promoted from the line jobs. The supervisor's recommendation is the basis for promotion, though seniority imposes some restrictions.

New inspectors are given brief instruction by the supervisor. This consists of a demonstration of the defects to be identified (some 18 different ones are to be judged for severity), and some instruction in how the panels can be handled, though there is no prescribed method. They are shown a copy of the written specification, which is in a book near the workplace. Since the specification is written in engineering terms, its main function for both workers and supervisors is to provide the names of the defects. The specification may describe defects of appearance, and they are also represented by limit samples.

The official limit samples are the sole property of the quality control department and are usually kept under lock and key. At the time of the experiment a cabinet of limit-sample equivalents at the far end of the work area was available to the supervisor.

Ordinarily, new workers are placed near more experienced inspectors whom they probably consult when in doubt, thus eliminating some of the follow-up essential to good training. They are instructed to confine their questions to the supervisor. He decides on the spot, consults the limit samples, or someone from quality control. Emphasis is always on identification of the defect. There is some reinspection of pieces passed, little or none of rejects. Several systems of identifying the work of individual inspectors have been tried without success, so there are no real records of inspector competence.

Most of the time, the typical inspector reaches a decision on his own, without reference to fellow workers, the supervisor, or the specification. Speed of handling and a manner of decisiveness are commonly considered characteristic of a good inspector.

Aside from the fact that there was no opportunity for consultation in cases where the inspector was in doubt, performance on the job-practice run appeared to be typical of work on the job.

[6] See footnote 4.

TABLE 4
MEAN SCALE VALUES ASSIGNED NEW SAMPLE (FACTORY STUDY)

Group	Panel No.									
	1	2	3	4	5	6	7	8	9	10
14 inspectors	6.3:	2.6	6.6	6.1	5.4	2.3	7.2	0.9	6.6	1.3
4 management men	7.2	1.2	6.0	5.8	4.5	3.0	8.2	1.5	6.2	1.2
"Exact" values, Customer Inspector	7	0	7	5	5	3	9	1	6	2

RESULTS

The Paired-Comparison Decisions

The ability of the industrial inspectors to distinguish between panels was estimated by noting the number of ranks resulting from the paired-comparison preferences, and by calculating individual and mean scale values for the panels as in the pilot study. Summary data are given in Table 4. The minimum number of ranks obtained from an individual subject was 8, with a mean of 9.1 ranks and a median of 9.6 ranks. Had the inspectors simply been sorting the good from the reject panels, there would have resulted—theoretically, at least—only two ranks, one for good panels, one for rejects.

More precise evidence on this point can be drawn from an examination of Table 5. This gives the critical ratios for the differences between mean scale values for all 45 pairs.

Section A of Table 5 shows that in the 20 pairs where a reject piece appeared with a good one, the critical ratios were all significant at the .01 level of confidence. All differences for the four pairs where a reject appeared with the limit sample were also significant at the .01 level, as shown in Section B. Section C shows that in six cases where reject panels were paired with each other, the critical ratios ranged in significance from .01 to .60. In Section D appear the five cases where a good panel was paired with the limit sample: here the ratio was significant at the .01 level in two cases, at the .10 level in one case, and at the .20 level in the remaining two cases. Only when the good panels appeared with each other did the significance of the differences fall near chance expectancy, ranging from the .10 level to no difference (see Section E). For all 45 pairs the differentiation is far above chance expectancy. Hence it seems that our industrial subjects did considerably more than merely dichotomize the sample.

The reliability of the mean scale values in this sample compared favorably with that of the samples in the pilot study. The rank-order coefficient was .94 ($r = .90$) for the Customer Inspector on two trials. Split-half reliability for random halves of the inspector group was rho = .86, $r = .90$, which become .92 and .98, respectively, when stepped up by the Spearman-Brown formula. When the inspector group was divided into matched halves, representing the same assortment of jobs in each half, the reliability figures were rho = .93, $r = .975$, which become .96 and .99 when stepped up.

Accuracy of the inspectors' decisions—their ability to differentiate the good from the reject pieces in the sample—may be estimated in several ways. Probably the most definitive evidence is given in Table 5 which shows that all reject panels were distinguished from the good ones and from the limit sample at the .01 level of confidence by the group as a whole. On this basis, accuracy of the group is 100%.

TABLE 5

DIFFERENTIATION BETWEEN PANELS, FACTORY STUDY

(Critical ratios for difference between mean scale values assigned by 14 inspectors.)

Pair	CR	p
A. Reject vs. Good Panels		
2–1	6.49	.01
2–3	8.51	.01
2–4	6.73	.01
2–7	7.07	.01
2–9	6.56	.01
6–1	8.88	.01
6–3	14.33	.01
6–4	10.00	.01
6–7	11.13	.01
6–9	8.60	.01
8–1	11.73	.01
8–3	17.81	.01
8–4	13.00	.01
8–7	11.07	.01
8–9	11.17	.01
10–1	8.77	.01
10–3	11.52	.01
10–4	9.23	.01
10–7	9.08	.01
10–9	7.61	.01
B. Reject vs. Limit Sample		
2–5	5.09	.01
6–5	7.38	.01
8–5	10.22	.01
10–5	7.45	.01
C. Reject vs. Reject Panels		
2–6	0.638	.60
2–8	3.46	.01
2–10	2.20	.05
6–8	4.24	.01
6–10	2.10	.05
8–10	0.81	.50
D. Good vs. Limit Sample		
1–5	1.69	.20
3–5	2.92	.01
4–5	1.49	.20
7–5	2.90	.01
9–5	2.06	.10
E. Good vs. Good Panels		
1–3	0.68	.60
1–4	0.4	.70
1–7	1.43	.20
1–9	0.5	.70
3–4	1.08	.30
3–7	1.11	.30
3–9	0.00	—
4–7	1.86	.10
4–9	0.91	.40
7–9	0.95	.40

The accuracy of individual inspectors may be estimated by examining their individual records, scoring as an error all instances where a reject is given a scale value higher than or equal to the lowest value assigned to a good panel or to the limit sample. By this definition, 10 inspectors had no errors, or were 100% accurate. The remaining four had one error apiece, or were 90% accurate. Average accuracy of the group was 97%.

Because it would be simpler to incorporate a comparison of each piece being inspected with the limit sample than to apply the principle of the full design to the present inspection method, the "accuracy" of decisions on the nine pairs per trial when each panel appeared with the limit sample was checked for each subject. The decision was scored as correct when the limit was preferred to a reject, and when the good panels (which had been differentiated from the limit sample significantly) were preferred to the limit sample. On that basis, accuracy was 88.5% on the good panels, 91% on the reject panels, and 90% on the full sample. These data are given in Table 6.

The coefficient of correlation may also be used as an index of accuracy since the scale values assigned by the Customer Inspector are 100% accurate by definition. They correlated with the mean scale values assigned by the inspectors, rho = .87, r = .91. Mean scale values assigned by him and the supervisors may be considered as an additional criterion,

TABLE 6

NUMBER OF CORRECT DECISIONS, 14 INSPECTORS

Subject No.	Job-Practice Run			Comparison with Limit		
	Good*	Reject	Total	Good**	Reject	Total
1	2	0	2	5	4	9
2	4	1	5	4	4	8
3	4	0	4	5	4	9
9	4	0	4	5	3	8
10	4	0	4	3	4	7
5	2	1	3	5	3	8
6	6	4	10	4	2	6
8	5	1	6	4	4	8
14	3	0	3	5	3	8
15	2	2	4	5	4	9
7	4	2	6	4	4	8
11	6	3	9	4	4	8
12	4	1	5	5	4	9
13	5	1	6	4	4	8
Total	55	16	71	62	51	113
%	65	29	51	88	91	90

* $N = 6$, includes Limit.
** $N = 5$, Limit excluded.

and their correlation with the inspectors' scale values was rho = .88, $r = .95$.

Of particular interest to the management was the possibility that inspectors would differ in their rankings, the differences being attributable to their regular inspection jobs. Rank-order correlations of the mean scale values assigned by job groups with those assigned by all others were: final bulb inspectors, .95; production check inspectors, .95; production parts inspectors, .93; and process inspectors, Blow Room, .92. No significant differences attributable to the type of inspection usually performed were noted. Correlations between other subgroups based on sex and length of service, calculated in the same manner, were similar, showing no differences attributable to these factors.

The Job-Practice Decisions

Individual inspectors scored from 2 to the maximum possible 10 correct decisions on the job-practice run, ranging from 20% to 100% accuracy, and averaging 51%. These data are also in Table 6.

The whole group achieved an accuracy of 65% on the good panels, 29% on the rejects, and 51% on the whole sample. Since the Customer Inspector had selected only panels showing defects which could not be improved by reclaiming, the large number of decisions to send a panel to be reclaimed were all scored as errors. Had a decision to pass or reject been forced in each case, presumably half the guesses would have been correct. With this allowance accuracy becomes 81% for good panels, 60% for rejects, and 73% for the whole sample.

SIGNIFICANCE OF THE FACTORY STUDY

Reliability of the Paired-Comparison Judgments

The factory study substantiated the findings of the laboratory investigation, showing comparable reliability, and in-

dicating that the method could be used to develop a reliable quality scale for defects of appearance.

Validity of the Paired-Comparison Judgments

The scale values developed by the industrial inspectors showed a satisfactory validity when correlated with the values assigned by the Customer Inspector. Further evidence of their validity may be obtained by correlating the mean scale values with the number of the inspectors' decisions to pass, reject, or reclaim each panel on the job-practice run. These rank-order correlations were .94 between scale value and number of decisions to pass the panel, −.81 between scale value and number of decisions to reject it, and −.71 between scale value and number of decisions to reclaim it.

Accuracy of the Paired-Comparison Judgments

The outstanding result of the factory study seemed to be the *superior accuracy of the paired-comparison judgments over those given on the job-practice run.* The data are summarized below.

The mean scale values differentiated all good from all rejects at the .01 level of confidence—accuracy 100%. Individual inspectors assigned higher values to all good panels than to any reject in 136 of a possible 140 instances—average accuracy 97%. In the nine judgments per inspector when the limit sample was compared with each of the other pieces, the inspectors were 90% accurate. In judging the panels according to job practice, they were only 51% accurate.

These data were derived from the full complement of the quality control department for one shift. Projection of the loss of product attributable to inspector error to the full shift would indicate a loss of hundreds of pieces per day. Hence a sizable investment in changing inspection method to improve accuracy should pay for itself in recovery of lost product.

According to the data cited above, if the inspection procedure were changed to the full paired-comparison design, a 90% improvement in accuracy can be predicted. This would be practical only in small-lot sampling inspection. But if the method were changed to require that all pieces be compared with a limit sample instead of with the subjective or "memory image" standard currently employed, accuracy might be increased by 76%. It seemed that this change could be made easily, even in 100% inspection.

A Debatable Point

Before recommending this change to management, it seemed wise to check one questionable point with further experimentation. Data on the accuracy of the comparisons with the limit sample were drawn from the responses made in the paired-comparison trials. Though it seemed reasonable to expect that similar decisions would be given if each piece were compared only with the limit sample, it was also possible that the subjects' decisions were favorably conditioned by the other 36 pairs in the sample. Accordingly, a plan to repeat only the comparisons with the limit sample was drawn up and the company so advised.

Changed Factory Conditions

The local management reported that soon after the experiment had been car-

ried out, a slump in the television industry had necessitated a six-month shutdown on the 20-inch television bulb. When production was resumed, not only had there been a 100% turnover in the inspection force and the work sample been lost despite precautions, but inspection methods had been changed considerably.

At that time the local management had not received any information from the experimenter, except a casual comment in a social conversation, that the paired-comparison decisions were apparently more accurate than those made on the job-practice trials, The Quality Control Director reported that "watching the experiment gave them some ideas which they incorporated in the inspection job when they had to set it up again for new people." He described the new method as follows:

Revised inspection method. The first step in preparing for resumption of production was the selection of a set of limit sample equivalents for each workplace. These were selected by the pooled judgment of the Quality Control Director, several of his experienced supervisors, and the Customer Inspector. The samples are stored near the workplace, and at the beginning of each shift the supervisor selects those samples which represent the particular defects appearing in current production and places them on a rack built over the conveyor belt by which the inspectors work. The limit then appears in the same physical relationship to the piece under inspection as it did in the experiment. (There is no change in the number of specific defects listed in the formal specification, but defects tend to occur in families, dependent on whether they are attributable to machine operation, temperature, quality of raw material, etc. Hence only two or three limit samples would be needed at a time.) The formal specification was simplified, illustrated with diagrams, and posted at the workplace.

Inspectors are hired or upgraded as before, and it is still the supervisor's responsibility to break them in. He usually begins his instruction by referring to the posted specification, but concentrates on the technique of comparing each piece with the limit sample which has been installed. The gist of his instruction is that pieces that are inferior to the sample are to be put in the hopper, while those that are as good as it is or better are to be placed elsewhere for transportation to final assembly or packing. The inspector still knows that he is separating good from reject pieces, but the emphasis has now been placed on the specific comparisons. The recording and reinspection systems are about as before. The supervior spends about the same amount of time in instruction, but he feels that the inspectors learn the job more quickly and maintain a more consistent performance than before.

It was obvious that job method had changed so much that further experimentation would not give results comparable to those already obtained. Therefore, information indicative of inspection efficiency under the two methods was requested from the regular production records. The experimenter suggested two items which might be pertinent: (*a*) The number of rejections at final inspection should indicate errors made by inspectors performing first or parts inspection; and (*b*) the accuracy of final inspectors should be indicated by the number of customer complaints (see footnote 5), since the pieces customers complain about are the ones for

TABLE 7

PRODUCTION RECORDS ON THE 20-INCH TELEVISION BULB

Item	Period A, First Method	Period B, Revised Method
Total panels produced by machine	213,828	219,108
No. rejects at first inspection	71,892	90,618
No. panels sent to assembly	141,936	128,490
No. rejects at final inspection	1,728	1,250
No. shipped to customer	16,715	33,059
No. rejects by customer	279	130

which the company actually makes replacements or refunds.

Accuracy of Inspection by the Two Methods

Production records on the 20-inch rectangular television bulb were reviewed for the six months preceding the shutdown (Period A) and for the six months immediately after production was resumed (Period B). The plant management supplied the data in Tables 7 and 8.

The critical ratios given in Table 8 for the differences in percentage of inspection losses by the two methods are all statistically significant at better than the .000001 level of confidence according to standard tables.

It will be noted that the machine efficiency was apparently better during Period A, a fact which refutes the possibility that the decreased losses due to inspection at later stages of manufacture can be attributed simply to better quality of production by the machine. Management men who assisted in the study report that demand was considerably higher during Period A than during Period B. During Period B the television set manufacturers were just recovering from a severe cutback themselves. As a result, their orders were smaller and their quality specifications higher. Therefore, the change in the percentage of customer rejects cannot be attributed to relaxing of the specification. Hence it seems reasonable to ascribe the improvement to the change in the method of inspection.

TABLE 8

PERCENTAGE COMPARISONS FOR INSPECTION LOSSES IN TWO PERIODS OF PRODUCTION (A AND B)

	% Rejects		Difference A minus B	CR	% Improvement B over A
	A	B			
First inspection	33.62	41.36	−7.74*	5.301*	−23.0*
Final inspection	1.217	0.973	.244	6.10	20.0
Customer rejects	1.67	0.39	1.28	12.8	76.8

* Indicates difference is in favor of Period A.

SUMMARY AND CONCLUSIONS

The results obtained in this study support the hypothesis stated in the Introduction that a methodology which is focused sharply on the actual judgments involved in an appearance inspection job might give data indicative of the basic inspection function and suggest conclusions which might be generalized to other inspection jobs.

The Method of Paired Comparisons

As an experimental technique, the method of paired comparisons seems to have the following advantages in a study of an industrial inspection job.

High Reliability

Reliability of the paired-comparison ranks of the work samples used in the study exceeded .90 for industrial subjects and two groups of untrained subjects. This is a considerable improvement over the reliability reported for most work-sample experiments in the literature, and a marked increase over the reliability of the criteria in the studies attempting to validate psychological tests. Not only were the scale values (or average ranks) consistent, but the relative rank assigned to part of a sample remained similar when that part appeared in a different sample. Reliability coefficients obtained in this study compare favorably with those reported for other uses of the method (12).

High Accuracy of Comparative Judgments

Accuracy of the decisions obtained in the paired comparisons exceeded that obtained by the traditional method in the job-practice run by at least 90%. Individual comparisons with the limit sample[7] were 76% more accurate than the job-practice decisions. The accuracy figures of 97% and 90%, respectively, are considerably higher than those reported for work-sample studies using other methods (9, 11, 13, 20).

Validity and Applicability of the Data

Validity coefficients of .91 and .95 for the mean scale values assigned by the inspectors with those assigned by the Customer Inspector and management, respectively, indicate that judgments obtained by the paired-comparison method are actually valid. Also, the mean scale values assigned by the inspectors correlated .94 with the number of their decisions to pass each panel on the job-practice run.

The major finding of this study is that specific or comparative judgments are more uniform and accurate than the type usually made on the job. The application of this finding by the company management in its revised inspection method resulted in an improvement of 76% in accuracy of final inspection. This is evidence of the practical applicability of the findings of the present study.

Possible Applications

The efficiency of the inspection process, as it is usually carried out, rests on the assumption that the inspector's idea or recollection of the limit sample is invariable whether the product is running mostly good or mostly bad. There is general recognition of the fact that quality of product changes frequently, since sampling inspection is customarily performed at 30-minute intervals, but there is little or no awareness of the effect of variations in product quality on inspector efficiency.

To the writer, the study suggests that, despite apparent aptitude, training, and experience, skilled industrial inspectors make their judgments in the same fashion as do untrained subjects—influenced to a considerable degree by the characteristics of the material under inspection at the time. The change in method suggested by the experiment apparently corrected this bias to a considerable extent, and was followed by a

[7] See footnote 4.

marked improvement in accuracy. Since the right of management to improve efficiency by selection and placement of present employees is subject to restriction or challenge under present industrial relations procedures, whereas control of work methods remains its prerogative, the advantage of data applicable to job method rather than limited to placement is obvious.

To 100% Inspection

The principle of requiring inspector judgments to be made against a limit sample may be applied to either sampling or to 100% inspection. Its application should result in a substantial increase in inspector accuracy. Writers, supervisors, and inspectors themselves agree that in the usual job situation the inspector's memory of the specification is continually being affected by the general quality level of the product he happens to be handling at the time. The presence of an invariable standard should eliminate or reduce this bias. In the event that the specification is changed, substitution of a new limit sample would change inspection judgments automatically, making retraining of inspectors unnecessary.

To Selection of Limit Samples

The method of the experiment should be useful as a scientific check on the adequacy of the specification or the limit sample which represents it. Frequently the description of a minimally acceptable unit of product is the result of considerable negotiation between product engineers during the writing of a sales contract and has little reference to the inspectors' ability to make the distinction in the degree of defect called for. Scaling several samples from normal production should indicate that point in the graded series where consistent distinctions can be made. Unless the specification describes a unit near the line of practical demarcation, some revision is indicated before the limit sample will actually function as intended.

The efficiency of any particular limit sample can be checked by using it in several samples—from the same production run, of course—to determine whether it is consistently scaled at the lower limit of the good pieces, and significantly above the rejects.

Usually there is but one limit sample per defect designated for an item, while the production situation requires that inspection be performed at several locations at the same time. The design of the study could be used to identify other units of product which are not distinguishable from the original limit sample and these equivalents could be used at all inspection stations.

To Sampling Inspection

The procedure of this study is directly applicable to sampling inspection where the size of the sample does not exceed ten pieces per half hour. Inclusion of a limit sample in the sample lot should make possible a quantitative estimate of the quality of the total lot. Inspection records obtained in this manner should be less subject to falsification or error than are the present records.

The method may be used to develop a quality scale for *degree* of defect for those defects now judged simply on an acceptance basis. Such a scale would make possible the use of $\bar{X}$ and R charts in many situations where the volume required by the p or percentage defective charts prohibits their use in product or process control (7). Such an application would be pertinent to the product studied in this experiment.

REFERENCES

1. AYERS, A. W. A comparison of certain visual factors with the efficiency of textile inspectors. *J. appl. Psychol.*, 1942, **26**, 812-827.
2. DRAKE, C. A. Inspectors are born that way. *Fact. Mgmt Maint.*, 1940, **98** (4), 44-45, 102, 104.
3. DRAKE, C. A. New developments in the selection of factory workers. *Prod. Ser. Amer. Mgmt. Ass.*, 1940, No. 127, 32-43.
4. FEINBERG, R., & COLEMAN, J. H. Vision tests for inspectors insure good placement. *Fact. Mgmt Maint.*, 1945, **103** (1), 106-110.
5. GHISELLI, E. E. Tests for the selection of inspector packers. *Psychol. Bull.*, 1941, **38**, 735. (Abstract)
6. GIESE, W. J., & SAGEN, H. E. Ampoule inspection. *Drug Cosmetic Ind.*, 1950, **66**, 518, 574-575.
7. GRANT, E. L. *Statistical quality control.* New York: McGraw-Hill, 1946.
8. KEPHART, N. C. Visual skills and labor turnover. *J. appl. Psychol.*, 1948, **32**, 51-55.
9. KEPHART, N. C., & WISSEL, J. Vision inspection performance research, Foundation Case No. 1020. Unpublished study, Purdue Univer., 1949.
10. KERR, W. A. Vision tests for precision workers at R.C.A. *Personnel Psychol.*, 1948, **1**, 65-66.
11. LAWSHE, C. H. An experimental study of the relative efficiency of two methods for inspecting ophthalmic lenses. Unpublished study, Purdue Univer., 1945.
12. LAWSHE, C. H., KEPHART, N. C., & MCCORMICK, E. J. The paired comparison technique for rating performance of industrial employees. *J. appl. Psychol.*, 1949, **33**, 69-77.
13. LAWSHE, C. H., JR., & TIFFIN, J. The accuracy of precision instrument measurement in industrial inspection. *J. appl. Psychol.*, 1945, **29**, 413-419.
14. MCMURRY, R. N., & JOHNSON, D. L. Development of instruments for selecting and placing factory employees. *Advanced Mgmt*, 1945, **10**, 113-120.
15. MAHER, H., & FIFE, ISABELLE E. A biological-pharmaceutical checker selection program. *J. appl. Psychol.*, 1947, **31**, 469-476.
16. MANN, IDA, & ARCHIBALD, D. A study of a selected group of women employed on extremely fine work. *Brit. Med. J.*, 1944, **1**, 387-390.
17. RUNDQUIST, E. A., & BITTNER, R. H. Validation of tests for glass bottle inspectors at Owens-Illinois Glass Co. cited by Bingham, W. E. in Great expectations. *Personnel Psychol.*, 1949, **2**, 398.
18. SARTAIN, A. Q. The use of certain standardized tests in the selection of inspectors in an aircraft factory. *J. consult. Psychol.*, 1945, **9**, 234-235.
19. SHUMAN, J. T. The value of aptitude tests for factory workers in the aircraft engine and propeller industries. *J. appl. Psychol.*, 1945, **29**, 156-160.
20. TIFFIN, J., & ROGERS, H. B. The selection and training of inspectors. *Personnel*, 1941, **18**, 14-31.

(Accepted for publication October 20, 1954)